Seminarska naloga – Celovita podatkovna analiza in optimizacija procesa

Namen seminarske naloge

V seminarski nalogi boste na izbrani javni bazi podatkov izvedli celoten proces podatkovnega raziskovanja, modeliranja in optimizacije, kot ste ga spoznavali tekom vaj. Vaš cilj je razumeti delovanje izbranega procesa, identificirati pomembne dejavnike, razviti najboljši napovedni model ter pripraviti priporočila za optimizacijo procesa na podlagi rezultatov modeliranja.

Nalogo izvajata **dva študenta**:

- **En primer mora biti regresijski** (odvisna spremenljivka je **numerična**).
- **En primer mora biti klasifikacijski** (odvisna spremenljivka je **binarna**).

Od točke 1.4. dalje torej naredite vse ločeno za a) regresijski primer in b) klasifikacijski primer.

Nalogo oddate v roku v spletno učilnico

1. PRVI DEL: Priprava podatkov in osnovna analiza

1.1 Izbor podatkov

- Izberite javno dostopno bazo podatkov.
- Zastavite namen in cilje analize.
- Opišite vse spremenljivke (ime, pomen, tip).

1.2 Pregled in čiščenje podatkov

- Preverite manjkajoče vrednosti, podvojene zapise, ekstremne vrednosti.
- Po potrebi izvedite nadomeščanje manjkajočih vrednosti in utemeljite postopek.
- Jasno utemeljite vse sprejete odločitve.

1.3 Deskriptivna statistika z grafi

Za vsako spremenljivko:

- pripravite osnovne statistike (mean ± SD ali mediana (Q1–Q3), min–max oz. n (%)),
- pripravite grafični prikaz (histogram, boxplot, barplot …),
- interpretirajte opažanja (porazdelitve, odstopanja, posebnosti).

Izpišite tabelo z opisno statistiko za vse spremenljivke.

1.4 Bivariatna analiza

- Preverite povezave med vsako spremenljivko in odvisno spremenljivko.
- Uporabite ustrezne statistične teste (korelacije, χ^2, t-test, Mann–Whitney, ANOVA …).
- Dodajte grafične prikaze.
- Interpretirajte rezultate v kontekstu izbranega procesa.
- Rezultate predstavite v skupni tabeli

1.5 Izbor spremenljivk (Feature Selection)

Uporabite eno ali več metod za izbiro pomembnih spremenljivk:

- random forest importance,
- LASSO / elastic net,
- RFE,..

Na koncu določite, katere spremenljivke boste vključili v modele.

2. DRUGI DEL: Gradnja in ocenjevanje modelov

2.1 Priprava podatkov za modeliranje

- Podatke razdelite na učno množico (80%) in testno množico (20%).
- Če so podatki časovno odvisni → zadnjih 20 % uporabite kot testno množico.

2.2 Gradnja modelov

- Zgradite vsaj 5 modelov s pomočjo različnih algoritmov.
- Eden med temi 5 modeli mora biti linearni ali logistični regresijski model.
- Pri vseh modelih nastavite hiperparametre (ne uporabljajte privzetih nastavitev).
- Uporabite 10-fold cross-validacijo na učni množici za testiranje uspešnosti algoritmov (pazite na seme).

2.3 Metrike napovedne uspešnosti

- Regresija: R^2, RMSE, MAE, MAPE.
- Klasifikacija: AUC, Accuracy, Sensitivity, Specificity, PPV, NPV, F1.
- Rezultate predstavite v tabeli (mean ± SD čez 10-fold CV) in grafično.

Tabela: Primerjava modelov (validacijska množica)

Model	Tip	Parametri	Metrike (mean±SD)	AIC/BIC	Komentar	Izbor

3. Izbor najboljših modelov in testiranje

- Izberite 3 najboljše modele glede na validacijske rezultate.
- Zgradite končne modele na celotni učni množici (80%).
- Preizkusite jih na testni množici.
- Izračunajte vse pomembne metrike in rezultate predstavite v tabeli in grafično.

Tabela: Primerjava modelov (validacijska množica)

Model	Tip	Parametri	Metrike	AIC/BIC	Komentar	Izbor

4. Interaktivna aplikacija (Simulacija in optimizacija)

Aplikacija mora omogočati:

4.1 Izbiro modela

- Uporabnik izbere enega izmed 3 najboljših modelov.

4.2 Napoved za posamezen ali skupinski vzorec

- vnos vrednosti spremenljivk (v polje, spreminjanje z drsnikom, izbor vrednosti…),
- izračun napovedi Y,
- grafični prikaz rezultatov.

4.3 Simulacija sprememb

Uporabnik lahko:

- spreminja vrednosti neodvisnih spremenljivk,
- opazuje **grafični prikaz vpliva spremembe posamezne neodvisne spremenljivke**,
- vidi **novo napovedano vrednost Y**,
- izvozi rezultate za izbrani vzorec v CSV.

To je ključni del optimizacije procesa. Aplikacija mora omogočati simulacijo optimizacije procesa.

5. Povzetek ugotovitev in priporočila za optimizacijo

Izberite najboljši model glede na rezultate na testni množici.

V zaključku opišite:

- katere spremenljivke imajo največji vpliv,
- ali je njihov vpliv pozitiven ali negativen,
- katere spremembe so procesno smiselne,
- kaj priporočate vodstvu.

Povzetek mora biti utemeljen in razumljiv vodstvu podjetja.

6. Testiranje globalnih sprememb (Simulacije)

- Iz izbranih modelov definirati **globalne spremembe** (npr. zmanjšanje časa, povečanje kakovosti).
- Pripravite novo različico testne množice z uvedenimi spremembami.
- Izračunajte napovedi pred in po spremembah.
- Primerjajte metrike in razložite učinke.

Tabela: Pred–potem

Model	Metrika	Pred	Po	Razlika	Interpretacija

7. Six Sigma analiza – PREJ in POTEM

- Izračunajte DPMO in sigma nivo pred spremembami.
  - regresijski: napake definirate preko tolerance,
  - klasifikacijski: napake = napačne klasifikacije.
- Izračunajte DPMO in sigma nivo po spremembah.
- Pripravite tabelo in analizo izboljšave procesa.

Dodajte kratko interpretacijo:

- kako se je proces izboljšal,

- ali je sprememba statistično in procesno pomembna,
- ali upravičuje implementacijo v praksi.

Tabela: Primerjava sigma stopnje

Model	DPMO PREJ	Sigma PREJ	DPMO POTEM	Sigma POTEM	Izboljšava

8. Končni povzetek (Executive Summary)

Pripravite ½–1 stran povzetka:

- kaj je bil problem,
- kateri modeli so najboljši in zakaj,
- katere spremenljivke je smiselno optimizirati,
- kakšni so učinki sprememb,
- kako se je sigma stopnja izboljšala,
- priporočilo za implementacijo sprememb.

9. Oddaja naloge

Oddati morate:

1. **Python Notebook (.ipynb)** ali **R skripto (.Rmd)** – z vso kodo in rezultati,
2. **Končno poročilo (.docx ali .pdf)** z razlago in grafi,
3. **Delujočo aplikacijo** (npr. ShinyApp, Streamlit ali lokalno izvedbo z navodili za zagon).